RESEARCH

A Conceptual Model for Describing Processes of Crop Improvement in Database Structures

Ian H. DeLacy, P. N. Fox, Graham McLaren, Richard Trethowan, and Jeffrey W. White*

ABSTRACT

Rising research costs, broadening goals, intellectual property rights, and other concerns increase the need for robust management of crop improvement data. The data model of the International Crop Information System (ICIS) allows breeding processes to be recorded unambiguously in a relational database. This paper describes this model, which underlies the Genealogical Management System (GMS) of ICIS. The model recognizes three classes of methods by which genetic material is advanced. Generative methods such as crossing or mutagenesis increase variation. Derivative methods usually involve selection, and maintenance methods conserve the genetic makeup of germplasm, such as in seed multiplications. Unlike systems that only track pedigrees, the model describes steps of selection. Applications are illustrated for self-pollinating, outcrossing, and clonally propagated crops. The ICIS GMS is in use for species including rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), potato (*Solanum tuberosum* L.), common bean (*Phaseolus vulgaris* L.), lesquerella [*Lesquerella fendleri* (Gray) S. Wats.], and witloof chicory (*Cichorium intybus* L.). The International Rice Information System, based on ICIS, holds more than 2.6 million unique identifiers for germplasm accessions, crosses, populations, and lines, requiring about 900 megabytes of storage space, which can easily be managed on a personal computer. The GMS model appears suited for widespread use in managing data on crop improvement.

I.H. DeLacy, Univ. of Queensland, School of Land, Crop and Food Sciences, Brisbane, Queensland 4072, Australia, and Australian Center for Plant Functional Genomics (ACPFG), Brisbane 4072, Australia; P.N. Fox, ACIAR, GPO Box 1571, Canberra ACT 2601, Australia; G. McLaren, Generation Challenge Program, CIMMYT, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; R. Trethowan, Univ. of Sydney, Plant Breeding Institute, PMB 11, Camden, NSW 2570, Australia; J.W. White, USALARC, USDA-ARS, 21881 N. Cardon Ln., Maricopa, AZ 85138. Received 21 Jan. 2009. *Corresponding author (jeffrey.white@ars.usda.gov).

Abbreviations: COP, coefficient of parentage; DMS, Data Management System; GID, germplasm identifier; GMS, Genealogical Management System; ICIS, International Crop Information System.

Increasing operating costs, broadening research objectives, concerns over intellectual property rights, and increasing flows of data on molecular markers and gene sequences are fueling demand for more efficient management of crop research data. Electronic databases permit managing large sets of data, and the Internet can facilitate data exchange over large distances. Furthermore, various software tools exist or can be envisioned that would improve routine operations such as preparation of seed packets and field books or assist breeders in decisions on crosses and selections. The decision support role includes improved modeling of genotype × environment effects through consideration of the distant and recent ancestry of the lines (Crossa et al., 2006), use of association mapping (Parisseaux and Bernardo, 2004; Crossa et al., 2007; Stich et al., 2008), and detection of genetic regions under selection (Jordan et al., 2005; Cane et al., 2008). These considerations suggest large benefits from providing researchers access to integrated crop information systems. Indeed there are many examples of partial

Published in Crop Sci. 49:2100–2112 (2009).
doi: 10.2135/cropsci2009.01.0020
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

implementation of such systems (e.g., White et al., 1990; Fox et al., 1997; Tinker and Deyl, 2005; van Berloo and Hutton, 2005; Gingle et al., 2006). To our knowledge, the most complete public sector example is the International Rice Information System (Bruskiewich et al., 2003; McLaren et al., 2005), which is based on the open source International Crop Information System (ICIS; www.icis.cgiar.org; verified 24 Aug. 2008).

Developing an integrated crop information system is a costly and potentially risky undertaking, making the task unattractive for individual crop research programs. Collaborative software development offers one approach for sharing costs and reducing risk. Among advantages of this strategy is that prototypes can be constructed and tested in a more dynamic and robust manner than is possible in isolated software projects. It also provides greater assurance that continuity can be maintained as individual breeding or research programs come and go in the face of shifting priorities and budgets (White et al., 2007). For crop research information systems, a basic step toward collaborative development is to establish a conceptual model for describing the diverse procedures and nomenclature used by different crop improvement programs. In developing ICIS, a single data model was sought that could manage genealogies of any crop. "Genealogy" was interpreted in its broad sense, including processes that create genetic variation such as crossing, mutagenesis, and formation of transgenic plants, as well as selection to reduce or focus variation in the development of new lines or cultivars and maintenance methods such as regeneration and seed increase that aim to keep genetic variation constant (Fox and Skovmand, 1996). Examples of such information are found in cultivar and germplasm registration descriptions published in journals such as *Journal of Plant Registrations* and *Canadian Journal of Plant Science*.

This paper presents the theoretical model developed for ICIS as the core of its Genealogical Management System (GMS) and illustrates application of the model with examples for the breeding of rice (*Oryza sativa* L.), a self-fertilizing species; maize (*Zea mays* L.), an outcrossing species; and potato (*Solanum tuberosum* L.), a clonally reproduced species. Our focus is strictly on the handling of genealogies in GMS. Other components of GMS manage such information as germplasm names, locations and dates of creation and release, associated bibliographic references, and intellectual property status. In ICIS, characterization and evaluation data are primarily held in a parallel data management structure, the ICIS Data Management System (DMS), described by McLaren et al. (2005). Throughout, we refer to "seed," but in most cases, concepts are directly applicable to other planting materials such as tubers, stem cuttings, and rhizomes, and the GMS model is currently in use not only for seed propagated species such as wheat (*Triticum aestivum* L.), rice, and maize but also for vegetatively propagated crops including potato, sweet potato [*Ipomoea batatas* (L.) Lam.], and taro [*Colocasia esculenta* (L.) Schott].

THE GMS DATA MODEL

To accommodate multiple crops and breeding programs, the GMS model handles data on genealogy and nomenclature for all crops and requires that a minimal set of crop-specific parameters or software be provided separately (e.g., through crop or user profile data or as subroutines). Furthermore, all specific instances of germplasm are uniquely identified. Specific instances of germplasm can be thought of as equivalent to samples of seed (or clones) that exist or existed at some time, regardless of whether they were produced by natural or managed processes.

The fundamental issue for design of the data model was how to describe the processes whereby seed associated with one identifier gives rise to subsequent samples of seed that require new identifier(s). These processes would typically involve the creation of new variation through genetic recombination by crossing, the reduction of variation through selection, or the maintenance of genetic variation through seed increase. However, they also could include methods such as mutagenesis, polyploidization, or genetic transformation.

In the ICIS GMS model, each instance of germplasm is identified with a unique number germplasm identifier or GERMPLASM_ID (GID). Thus a cross, subsequent generations of populations or selections, and resulting lines all receive different GIDs. The basic concept is "if you wouldn't mix the packets of seed, grains, or clones then they each have a separate identifier (GID)." As explained subsequently, the GID also serves to identify the groups of germplasm that have originated from a significant event such as crossing, the immediate source material (plants of the preceding generation), and the founding stock for seed multiplications.

All processes used to modify germplasm over cycles of crossing, selection, or propagation are assigned to a specific "method." A critical distinction in the GMS model is whether a given method is intended to (i) increase, (ii) reduce or repartition, or (iii) maintain genetic variation as measured by allelic or gametic diversity. These three options lead to classifying methods as generative, derivative, or maintenance, respectively. In crop improvement, genetic consequences of these methods depend on the reproductive system of a given crop (Table 1), the genetic structure of the populations being manipulated, and the breeding strategies being used to achieve genetic change.

Generative Methods

Generative methods are intended to increase allelic diversity by combining alleles from different progenitors through crossing or mutating genes through mutagenesis, introducing new genes through transformation

Table 1. Reproductive (breeding) systems for crops considered for the Genealogical Management System (GMS) model.

Reproductive system	Descriptions and comments	Examples
Self-pollinated (self-fertilized)		Wheat, rice, barley, soybean, common bean
Predominantly self-pollinated		Cotton, pigeonpea, canola
Open-pollinated	Includes cross-pollinated or cross-fertilized	Maize, pearl millet, sorghum, cucurbits
Self incompatible or dioecious		Rye, white clover, papaya
Vegetatively or clonally propagated	Includes clonal propagation. For sexual generations, the generative, derivative, and most maintenance methods will usually be identical either to self or cross fertilized crops	Potato, cassava, yam, taro, sugar cane, pineapple, strawberry
Apomictic	Seed produced by asexual means. Catered for largely by methods for vegetatively propagated crops.	Green panicgrass, buffelgrass

or combining whole genomes through polyploidization. Generative methods based on crossing are conveniently classified by sources of female and male gametes. The source of female gametes is assumed to be determined by the source of the seed, which can be a single plant, a selected set of plants, or a random set of plants. Similarly, the source of the pollen can be from the same plant(s) as the female source, another single plant, a selected set of plants, or a random set of plants. These combinations define an array of possibilities with a dimension of three (female sources) by four (male sources) that represent the possible crossing practices used in crop improvement. Table 2 relates female and male sources of gametes to terminology more common to crop improvement such as "three-way cross" and "population backcross."

Derivative Methods

Derivative methods are processes applied to a single source of seed and are designed to reduce or repartition genetic variation (Table 3). Example methods are self-fertilization of lines in segregating populations, which reduces allelic diversity through inbreeding (in turn increasing homozygosity), production of double haploid lines, or randomly mating selected plants within a population. All instances of germplasm produced by derivative or maintenance methods from the same generative source form a related group of germplasm. Each derivative in such a group is linked to the GID of the generative source, the group GID. It is also linked to its immediate source via the source GID. For example, a sample of F_3 seeds is linked to its F_1 group and to the F_2 from which it was derived.

Maintenance Methods

Maintenance methods, again applied to a single source of seed, represent deliberate attempts to maintain a specific level of genetic variation with the objective of creating new instances of germplasm that are as similar to the source germplasm as possible (Table 4). Common examples would be methods used for increases of germplasm accessions, genetic stocks, or foundation seed. Besides the GID of the germplasm, identifiers for the group and the founder would link to the record (Fig. 1).

Examples of Methods

Examples of methods are given in Table 5, and the full set of methods is available on the ICIS wiki under "Ontologies and Controlled Vocabularies" (ICIS, 2008a). Method types are specified to distinguish between generative (GEN), derivative (DER), and maintenance (MAN) methods, respectively. Method groups are used to identify the relevant breeding system. Method group G indicates generic methods applicable to any breeding system; S indicates methods appropriate for naturally self-pollinating species; O, for open-pollinating species; and C, for methods of clonal propagation. The method code is a short mnemonic for the method (e.g., "C3W" for a three-way cross). Each method has a name that also indicates the type where necessary: CF for cross-fertilizing species, CP for clonal propagation, and SF for self-fertilizing species. Each method has a description and a number of allowable progenitors, Total progenitors. Allowable values are N for a variable number, 2 for two progenitors, 1 for a single progenitor, and −1 if the method is not a generative method and hence allows only one source germplasm and identification of a group germplasm.

The methods supplied in GMS are a compromise between a parsimonious set based on the model applied in Tables 2, 3, and 4 and an attempt to define every possible method that could be used in crop improvement. One example of this compromise occurs between the export and import protocols and the acquisition, seed increase, and cultivar generation protocols. Only one generic method is given for methods for import and export of genetic materials, but a series of detailed methods specifying the type of material acquired, increased, or released as cultivars are given for the latter protocols. The richness and specificity of the method ontology can be expanded in two ways. First, new methods can be added as needed, but the benefits of exact descriptors have to be weighed against the loss of benefits of data standardization. Second, notes or "method attributes" can be attached to specific instances of the use of methods. For example, if it is important to note the number of plants harvested in a bulk, this number can be attached as an attribute of the method bulk in the ICIS. These attributes can be formatted and hence are amenable to computer processing.

Table 2. Outline of a classification of generative methods based on sources of female and male gametes.

Female source	Male source	Generative methods
Single plant	Same single plant	Selfing
	Different single plant	Single cross
		Three-way cross
		Double cross
		Full diallel cross
		Full diallel cross bulked
		Half diallel cross
		Half diallel cross bulked
		Partial diallel cross
		Partial diallel cross bulked
		Backcross
		Backcross recessive
		Interspecific cross
		Narrow based tester, line CF
	Selected bulk	Female complex top cross
		Selected pollen cross
		Narrow-based tester, line CF
		Broad-based tester, Line CF
		Narrow-based tester, POP CF
	Random bulk	Female complex top cross
		Random pollen cross SF
Selected bulk	Single plant	Male complex top cross
		Backcross recessive
		Narrow-based tester line CF
	Same selected bulk	Polycross
		Random mating
	Different selected bulk	Complex cross
		Gametocide-mediated OP SF
		Male sterile–mediated OP SF
		Hand-mediated OP SF
		Full diallel cross
Selected bulk	Different selected bulk	Full diallel cross bulked
		Half diallel cross
		Half diallel cross bulked
		Partial diallel cross
		Partial diallel cross bulked
		Subset cross
		Population backcross
		Interspecific cross
		Selected pollen cross CF
		Narrow-based tester, POP CF
		Broad-based tester, POP CF
	Random bulk	Random pollen cross CF
		Population cross
Random bulk	Single plant	Male complex top cross
		Population backcross
		Population backcross recessive
	Selected bulk	Population backcross
		Population backcross recessive
		Selected pollen cross CF
		Broad-based tester, POP CF
		Population cross CF
	Same random bulk	Open pollination
	Different random bulk	Complex cross
		Gametocide-mediated OP SF
		Male sterile–mediated OP SF
		Hand-mediated OP SF
		Interspecific cross CF
		Random pollen cross CF
		Open pollination CF
		Population cross CF

Representation of Genealogies

Pedigrees are chains of GIDs linked by methods describing the processes used to create each germplasm instance. Thus the F_1 seed from a cross is linked to the GIDs of the parents, or progenitors, and identified by a new GID (generative germplasm in Fig. 1). Multiple female or male parents are also readily accommodated. Parents with unknown identity can also be indicated with a method for historical pedigrees with incomplete information (Table 5).

For derivative germplasm (derivative germplasm in Fig. 1), the model tracks both the immediate source germplasm (e.g., the F_2 population or plant from which an F_3 selection was made, derivative source in Fig. 1) and through the group ID. The group ID identifies the initial germplasm that represents the last generative method from which the derivative germplasm was obtained and is simply the GID of the initial germplasm. The immediate source is itself an example of derivative germplasm unless the line itself is produced directly from the group as in the case of F_2 seed, for example. If records for selections are incomplete, the source may be unidentified even when the group germplasm is known.

Maintenance methods are similar to derivative methods and lead to instances of maintained germplasm (as in Fig. 1). The group (original cross) of maintained germplasm such as from land races or heirloom cultivars is often unknown. The state of inbreeding of such germplasm often needs to be assumed from the natural reproductive behavior of the species. For example, collected rice samples are assumed homozygous, while collected maize samples are assumed heterozygous. Such germplasm with unknown origin can appear as the source or indeed the group of derivatives or maintenance descendants.

A third relationship is often needed for managing maintained germplasm. This is the founding sample (Fig. 1). For example, when a sample is received by a genebank (by import or collection), this becomes the founding sample for an accession, and all subsequent regenerations of that sample form

Table 3. Outline of a classification of derivative methods based on sources of female and male gametes in relation to derivative methods.

Female source	Male source	Derivative methods
Single plant	Same single plant	Single seed descent
		Plant identification
		Single plant selection
		Restorer selection
		Double haploid production
	Selected bulk	Single plant selection
		Selected bulk
		Double haploid production
		Mass selection
	Random bulk	Single plant selection
		Selected bulk
		Double haploid production
		Mass selection
Selected bulk	Same selected bulk	Purification
		Rouging
		Selected bulk
		Full mass selection
	Random bulk	Half mass selection
Random bulk	Same bulk	Random bulk

Table 4. Outline of a classification of methods for maintaining germplasm based on the reproductive system of the crop.

Reproductive system	Examples of maintenance methods
Self-pollinated	Seed increase from a single seed
	Seed increase from a single spike or pod
	Seed increase from a single plant
	Seed increase from a number of selected plants
	Forming a pure line
	Production of breeder's seed
Open-pollinated	Open pollination of unselected individuals in a isolated area, and all seed bulked
	Elimination of off types from an open-pollinated population
	Seed increased through a full diallel cross
	Seed increased through a partial diallel cross
	Forming a hybrid cultivar
	Forming a synthetic cultivar
Vegetatively reproduced	Propagation through normal vegetative material
	Propagation of a clone via tissue culture
	Regeneration of a clone via tissue culture
	Maintenance of a clone in a field collection
	Formation of a clone as a cultivar

part of the accession and inherit the passport data associated with the founding sample. This is a management relationship rather than a biological relationship and often carries obligations of stewardship and intellectual property rights.

Use of the GIS in Data Processing

Pedigrees constructed through linked GIDs are unambiguous and free of the naming conventions that have traditionally been used to record pedigrees and are subject to recording errors. Pedigrees recorded using the GMS model are also readily computable in the sense that it is easy to calculate coefficients of parentage between germplasm samples and other measures of genetic relationship and contribution.

The GID also provides a scaffold for annotating germplasm with passport, characterization, and evaluation data. The relationships, when properly managed, provide an opportunity for identifying sets or neighborhoods of germplasm. For example a germplasm accession in a genebank is actually a set of samples related to the founding sample by maintenance methods. A maintenance neighborhood of a particular

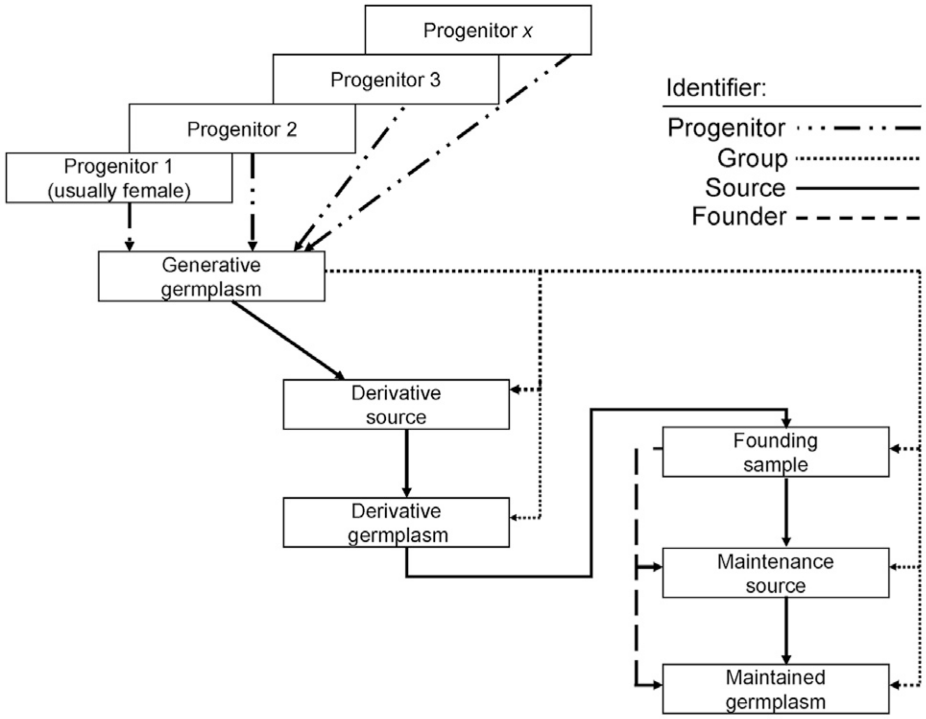

Figure 1. Diagrammatic representation of the different types of relationship among germplasm used in the Genealogical Management System (GMS) model. Progenitors are used in a generative process (e.g., crossing) to give rise to generative germplasm. This material then undergoes selection and or selfing through derivative processes. A germplasm sample from anywhere in this sequence may give rise to a founding sample in a collection, which may then be maintained through one or more seed multiplications. All germplasm that is derived or maintained from the same generative germplasm shares a common group identifier, and all samples maintained in specialized collections, such as in a germplasm bank or genetic stock collection, may be linked to the founding sample through the founder identifier.

Table 5. Examples of methods used to generate germplasm defined in the International Crop Information System (ICIS) Genealogical Management System (GMS). See ICIS (2008a) for a complete enumeration of methods of germplasm creation.

Type	Group	Code	Name	Description	Total progenitors
Methods for storing historical pedigree data with incomplete information					
GEN	O	PGM	Unknown generative method CF	Unknown generative method for storing historic pedigrees for cross-fertilizing species.	N
DER	O	PDM	Unknown derivative method CF	Unknown derivative method in cross fertilized species: for storing historic pedigrees	–1
DER	O	GMS	Generic mass selection CF	Production of next generation by selecting either before or after pollination, but where exact method is unknown. Primarily used for historic records.	–1
GEN	S	UGM	Unknown generative method SF	Unknown generative method for storing historic pedigrees for self-fertilizing species.	N
GEN	S	BDU	F_1 Backcross, cytoplasm unknown SF	Cross of F_1 to recurrent parent when the direction of the cross is unknown for storing historic pedigrees for self-fertilizing species.	2
GEN	S	BRU	F2 Backcross, cytoplasm unknown SF	Cross of F2 to recurrent parent when the direction of the cross is unknown for storing historic pedigrees for self-fertilizing species.	2
DER	S	UDM	Unknown derivative method SF	Unknown derivative method in self-fertilizing species: for storing historic pedigrees	–1
DER	C	CDM	Unknown derivative method CP	Unknown derivative method in clonally propagated species	–1
Generic maintenance methods					
MAN	G	IDN	Plant identification	Identifying and naming a plant or population.	
MAN	G	NSI	Seed increase	Increase seed of a cultivar, line, population or accession.	–1
MAN	G	ISE	Import	Import seed, clones or tissue culture of a cultivar, line, population or accession.	–1
MAN	G	SSL	Store seed long term	Store seed of a cultivar, line, population or accession. Genetic drift is expected.	–1
MAN	G	VFS	Foundation seed	Producing Foundation seed. Pure seed derived from Breeders seed (usually kept by seed producing organization)	–1
MAN	G	VCS	Certified seed	Producing Certified seed. Pure seed produced under supervision by government protocols.	–1
MAN	G	VCR	Cultivar release	Release a cultivar	–1
Generative methods for inbreeding crops					
GEN	S	C2W	Single cross	Cross between two single plants. If both parents are fixed (pure) inbred lines there will be no segregation for gametes or genotypes in the F_1 and theoretically all crosses will result in the same genetic outcome. In plant breeding practice the theoretical situation is rarely encountered. In spite of this the usual practice is to bulk the seed. However, in genetics studies it is often necessary to keep individual seed separate. When this is done, a separate entry in the germplasm table is required for each entity (seed) kept separate.	2
GEN	S	C3W	Three-way cross	Cross between two plants, one an inbred line and one a single cross (usually an F_1) and thus segregating for gametes. In the theoretical case, rarely achieved, the inbred line would be fixed and the F_1 a cross between fixed lines. The segregation for gametes results in different genetic outcomes among different progeny, hence a number of crosses using the same F_1 is usually made. Since different F_1s are genetically the same (theoretically) only one F_1 is required. In plant breeding programs the different crosses are usually bulked. Again, if individual seeds are kept separate a different entry is required in the germplasm table.	2
GEN	S	BC	Backcross	Backcross to recover a specific gene. The coding in the genealogical table records which parent was used as the female in each cycle. A different entry is required in the germplasm table for each entity kept separate.	2
GEN	S	BCR	Backcross recessive	Backcross to recover a recessive gene. As this requires a self-fertilization (derivative method) in the process some ICIS administrators may distinguish this as a separate method. A different entry is required in the germplasm table for each entity kept separate.	2
GEN	S	CIS	Interspecific cross	Cross between two species, usually requiring embryo rescue. The problem with making this a separate method is that the species cross could be made by any of the previous (101–108) or following (110–113) methods.	2
GEN	S	CSP	Selected pollen cross SF	A bulk of pollen from a selected set of males used to pollinate a female inbred line.	N
GEN	S	CGO	Gametocide-mediated OP SF	Open pollination in a self-fertilized species achieved through the use of a male gametocide.	N

Table 5. Continued.

Type	Group	Code	Name	Description	Total progenitors
GEN	S	CMO	Male sterile–mediated OP SF	Open pollination in a self-fertilized species achieved through the use of a dominant male sterile gene.	N
GEN	S	MIP	Induced mutation population SF	A population derived from inducing mutation in a inbred line.	1
GEN	S	C2WL	Somoclone SF	Variation induced through tissue culture of a inbred line.	1
GEN	S	ALP	Allopolyploid SF	Polyploid formed by doubling the chromosomes of a cross between two or more species. Wheat is an allopolyploid as it contains genomes from three different species.	1
GEN	S	AUP	Autopolyploid SF	Polyploid formed by doubling the chromosome number of a species. Lucerne (alfalfa) is an autopolyploid with 4 sets of the same genome.	1
GEN	S	HAP	Haploid SF	Individual with chromosome content of reduced gamete. Often formed by female progenitors crossed with a haploid inducer.	1
GEN	S	TRN	Transgenic nucleus SF	Individual derived from genetic transformation of the nucleus in a self fertilizing species.	1
GEN	S	TRC	Transgenic cytoplasm SF	Individual derived from genetic transformation of a cytoplasm inclusion (e.g., chloroplast) in a self-fertilizing species.	1
Derivative methods for inbreeding crops					
DER	S	MIL	Induced mutation line	A recognized mutation selected from an induced mutation in a line of a self-fertilized species.	–1
DER	S	DDH	Double haploid line	Individual produced by doubling haploid individual usually by anther culture in a self-fertilized crop.	–1
DER	S	DSP	Single plant selection SF	Derivation through selection of a single plant, inflorescence, fruit or seed from a self-fertilizing population.	–1
DER	S	DSB	Selected bulk SF	Derivation through bulking seed from a selected set of single plants from a self-fertilizing population.	–1
DER	S	DRB	Random bulk SF	Derivation through bulking seed from a random selection of single plants from a self-fertilizing population.	–1
DER	S	DSD	Single seed descent SF	Derived through the production of a single individual without selection from each individual in a segregating population.	–1
Maintenance methods for inbreeding crops					
MAN	S	NSP	Seed increase plant SF	Seed increase from a single seed, spike, pod, or plant in a self-fertilized species.	–1
MAN	S	NMX	Seed increase mixture SF	Seed increase from a number of selected plants in a self-fertilized species.	–1
MAN	S	NBK	Seed increase bulk SF	Seed increase from an unselected bulk in a self-fertilizing species.	–1
MAN	S	VPL	Pure line formation	Forming a pure line CV in a self-fertilizing species.	
MAN	S	VHY	Hybrid formation SF	Forming a hybrid CV in a self-fertilizing crop.	–1
MAN	S	VML	Multi-line formation SF	Forming a multi-line CV in a self-fertilizing crop	–1
MAN	S	VBS	Breeders seed production SF	Producing Breeders seed. Pure seed produced by breeder (usually some kept by breeder) in a self fertilizing crop.	–1
Generative methods for outcrossing crops					
GEN	O	P2W	Single cross CF	Cross between two single heterozygous plants derived from an open-pollinated population.	2
GEN	O	PFD	Full diallel cross CF	Each parent mated to all others, including all reciprocals but not selfs, usually not in isolation and all full sib and reciprocal families kept separate.	2
GEN	O	PFB	Full diallel cross bulked CF	Each parent mated to all others, including all reciprocals but not selfs, usually not in isolation and all seed bulked.	2
GEN	O	PHD	Half diallel cross CF	Each parent mated to all others, no selfs and reciprocals not recorded, usually not in isolation and full sib families kept separate.	2
GEN	O	PRM	Population random mating CF	Open pollination of a selected set of individuals in isolation and all seed bulked.	N
GEN	O	PBC	Population backcross CF	Backcross to introgress a gene into a population.	2
GEN	O	PBR	Backcross recessive CF	Backcross to introgress a recessive gene into a population.	2
GEN	O	PIS	Interspecific cross CF	Cross between two species.	2
GEN	O	PSP	Selected pollen cross CF	A bulk of stored pollen from a selected set of males used to pollinate a female population or plant.	N
GEN	O	PRP	Random pollen cross CF	A random bulk of stored pollen from some population used to pollinate a female population or plant.	2

Table 5. Continued.

Type	Group	Code	Name	Description	Total progenitors
GEN	O	TNL	Narrow-based tester, line CF	Test (Top) cross between a known plant and a narrow-based (1 or few plants) population. For practical reasons the tester population or line is used as the male which can be stored pollen. If the tester line or population is female this should be flagged as an attribute of the method.	N
GEN	O	TBL	Broad-based tester, line CF	Test (Top) cross between a known plant and a broad-based (many plants) tester. For practical reasons the tester population is used as the male which can be stored pollen. If the tester line or population is female this will be indicated by the position of the GID for the tester population.	2
GEN	O	PPO	Open pollination CF	Open pollination of an unselected set of individuals in isolation and all seed bulked.	1
GEN	O	PCR	Population cross CF	Cross between two populations.	2
GEN	O	PCC	Convergent cross	Series of single crosses, each cross then combined into double crosses, each of these then crossed etc.	2
Derivative methods for outcrossing crops					
DER	O	SLF	Self-fertilization CF	Self-fertilization of a plant or plants in a population.	–1
DER	O	DSO	Single plant selection CF	Selection of a single plant, inflorescence, fruit, or seed from a cross-fertilizing population.	–1
DER	O	PRS	Restorer selection	Restorer lines selected at the end of a program to back cross a gene which restores male fertility to lines carrying a male sterile cytoplasm (CMS).	–1
DER	O	FMS	Full mass selection	Production of next generation with selection before pollination, selecting on both male and female sides.	–1
DER	O	HMS	Half mass selection	Production of next generation with selection after pollination; selection on female side only.	–1
Maintenance methods for outcrossing crops					
MAN	O	MPO	Seed increase—open pollination CF	Open pollination of an unselected set of individuals in isolation and all seed bulked. Here the aim is to maintain a population, not recombine a set of selected families.	–1
MAN	O	MFB	Seed increase—full diallel cross bulked	Each parent mated to all others, including all reciprocals but not selfs, usually not in isolation and all seed bulked. The aim is to maintain a population, not recombine selected families.	–1
Maintenance methods for clonally propagated crops					
MAN	C	NCI	Clone maintained in the field	Clone maintained in a germplasm garden in the field in the traditional manner.	–1
MAN	C	NCT	Clone maintained through tissue culture	Clone maintained as a tissue culture.	–1
MAN	C	VCF	Clone formation	Formation of a clone as a cultivar.	–1

GID is the set of all germplasm related to it by maintenance methods. (Note that the maintenance neighborhood of a sample in a genebank is often larger than the accession since it may reach beyond the founding sample to other samples maintained by the donor, for example.) Users often want data on all germplasm in a neighborhood, not just on the sample identified; for example, the seed stocks available for all germplasm in a maintenance neighborhood. Similarly, derivative neighborhoods can be defined as the set of all germplasm related by derivative methods—sister lines, for example, or all lines tracing to a single plant or within a specified number of derivative generations. Plant breeders, plant scientists, and germplasm curators often want to see data on all samples in such neighborhoods when making breeding decisions. Such queries are facilitated by the GMS model.

Representing Incomplete Genealogies

In dealing with historical data, records may be incomplete or contain errors that prove impossible to correct. As mentioned above, various methods are available for describing individual materials that lack additional genealogical information, but the GMS model also accommodates for missing relations among germplasm. The key to this capability lies in the dual links between the source and group. The group (cross) is often known when the source is not. Thus, it is possible to have the group ID known (not zero) where the source ID is unknown. It is not allowed to know the source but not the group. However, there are many cases (e.g., when germplasm is selected from landraces) where the group as a cross is unknown and most likely, the parents never will be known. In these cases, the group may be an unknown derivative representing the "founding germplasm," which must have unknown source and group if it is a self-fertilizing species or unknown parents if it is a cross-fertilizing species. In calculating coefficients of parentage (COPs; Wright, 1922; Cox et al., 1986; Souza et al., 1998) from pedigrees with unknown links, certain assumptions need to be made about the degree of inbreeding, and these also depend on the prevailing breeding system.

EXAMPLE APPLICATIONS

Details of implementation of the GMS model are illustrated below by describing the development of a famous rice cultivar and recording the pedigree of a clonally reproduced potato cultivar. Examples for development of a maize inbred line and for data handling in a wheat recurrent selection scheme are provided in Supplementary Fig. S1. More detailed descriptions of these and other genealogies are given in the ICIS GMS documentation, which is available on the ICIS wiki under "TDM GMS Overview" (ICIS, 2008b).

Genealogy of IR 64

The development of the rice cultivar IR 64 at the International Rice Research Institute (IRRI) illustrates a straightforward case of crossing, selfing, and selection in an autogamous crop (Khush and Virk, 2005). The example includes handling of missing information and crossing to a wild species.

The representation of IR 64's genealogy in the ICIS GMS is outlined in Table 6. As an arbitrary starting point, we assume that the following materials, identified in Table 6, are known and already have GIDs assigned in the database: IR 262-43-8-1 (GID 36), GAM PAI 30-12-15 (GID 37), IR 1737 (GID 68), IR 1833 (GID 71), IR 773 A1-36-2-1 (GID 85), IR 773 A1-36-2-1*3/O NIVARA (GID 90), IR 2006 (GID 91), IR 1561-149-1 (GID 93), and BPI 121-407 (GID 95). A value of "p" in the columns of Table 6 for the progenitors or sources of these germplasm indicates that in the complete database, the GIDs would be provided from previously entered records.

The earliest cross considered is a single cross (code C2W in Table 5) between materials IR 263-43-8-1 (GID 36) and GAM PAI 30-12-15 (GID 37) to produce the F_1 IR 833 (GID 38). This was subject to single plant selection (DSP) for four generations to produce GIDs 97, 98, 99, and finally, 100, which is the line IR 833-6-2-1-1. The group germplasm for all these lines is IR 833 (GID 38).

The next cross considered is between IR1561-149-1 (GID 93) and IR 1737(GID 68), which gives rise to IR 2040 (GID 94). IR 1737 is the result of backcrossing and, assuming a high level of backcrossing, will be highly inbred so the method used to produce IR 2040 is again considered a single cross (C2W).

The next cross, IR2055 (GID 96), involves the female parent BPI 121-407 (GID 95), which is a gene bank accession with no additional genealogy information. Hence, it is registered as the result of an unknown derivative method (UDM). The male parent of IR2055 is an F_1 from the cross IR1833 (GID 71), as seen from its method code, C2W, meaning single cross. Thus, the new cross IR2055 is recorded as a three-way cross (C3W) with immediate parents GID 95 and GID 71.

The female parent of cross IR2061 is IR833-6-2-1-1 (GID 100), which is derived as described above. The male parent is the F_1 of cross IR2040, which was entered as GID 94. Thus, the final step in creating cross IR2061 is a three-way cross (method C3W), identified as GID 101.

IR2146 (GID 102) is the cross of IR773A1-36-2-1 (GID 85) onto IR 773 A 1-36-2-1*3/O NIVARA (GID 90), a third backcross of IR773A1-36-2-1 (GID 85) onto an accession of the wild species *O. nivara* S.D. Sharma and Shastry. IR773A1-36-2-1 is the recurrent parent, so IR2146 is the fourth backcross entered with method BC.

IR5236 (GID 108) is a single cross between two lines derived from crosses recorded as GID 91 and GID 102. The selection and derivation histories of the lines are recorded through GIDs 103 to 107. Representation of cross IR5338 (GID 114) (IR 2061-456-1-4 × IR 2055-475-2) is analogous to IR5236. Next, the double cross IR5657 (GID 115) was made between the last two F_1s (method code C4W), and a series of derivative lines were produced (GIDs 116 to 118).

IR18348 is a single cross between the lines IR5657-33-2-1 (GID 118) and IR2061-465-1-5-5 (GID 120). These are derivatives from the double cross just entered (IR5657, GID 115) and the cross IR2061 (GID 101). The male parent is a sister line to the parent IR 2061-456-1-4 (GID 111), already used for cross IR 5338. The two lines diverge from a common line source in the F_3 generation, IR 2061-456-1 (GID 110).

The last step in the development of the variety IR 64 consists of three cycles of single plant selection from IR18348 (GID 121), which are identified by GIDs 122 to 124. Finally, to indicate formal release as IR 64, GID 125 is recorded with the maintenance method "cultivar release" (method VCR).

A graphic representation of the pedigree of IR 64 is shown in Fig. 2. Only direct parents are shown in the tree. F_1s which are not direct parents and intermediate derivatives are omitted to keep the diagram simple.

Genealogy of a Potato Cultivar Including Clonal Propagation

The second example shows how a simple genealogy of a cultivar of a clonally reproduced species, potato, can be recorded in GMS (Table 7). Note that potato readily cross-pollinates, so methods for outcrossing species are used in recording portions of the genealogy. The example is for the cultivar Tarago (Kirkham and Wilson, 1984; Kirkham, 1999). It is assumed that the "founder" breeding clones, the cultivars Orion, Katahdin, and Catriona and the breeding clone V28-12, are already entered in GMS (GIDs 301 to 304). In this example, all the materials used as parents are recorded as being the result of normal vegetative propagation (method NCI). The first two-way cross in the pedigree is between Katahdin and Catriona (method P2W) and has GID 401. A clone, GID 421, selected by an unknown derivative method (method CDM), is recorded under its breeder identifier. This clone is released as the cultivar Glen Iliam (GID 431, method VCF) and is regenerated through vegetative propagation (method NCI) to produce GID 441, used in the cross V28_12/Glen Ilam (GID 501). The derived clone 60_7_2 is subsequently used for the cross Orion/60_7_2, from which the cultivar Tarago is ultimately obtained.

Table 6. Application of the International Crop Information System (ICIS) Genealogical Management System (GMS) model to breeding of the rice variety IR 64. Method codes are as defined in Table 5. Numbers under female parent, male parent, derivative group, and source of derivative group are germplasm identifiers. A value of “p” is used to indicate that in the complete database, an actual GID number would be provided. Cells containing a period (“.”) would have a null value in the database.

Germplasm identifier	Method code	Female parent	Male parent	Derivative group	Derivative source	Breeder’s identification
36	DSP	.	.	p	p	IR 262-43-8-11
37	ISE	.	.	p	p	GAM PAI 30-12-15
38	C2W	36	37	.	.	IR833
68	BC	p	p	.	.	IR1737
71	C2W	p	p	.	.	IR1833
85	DSP	.	.	p	p	IR773A1-36-2-1
90	BC	85	p	.	.	IR773A1-36-2-1*3/O NIVARA
91	UDM	p	p	.	.	IR2006
93	DSP	.	.	p	p	IR1561-149-1
94	C2W	93	68	.	.	IR2040
95	UDM	.	.	p	p	BPI 121-407
96	C3W	95	71	.	.	IR2055
97	DSP	.	.	38	38	IR833-6
98	DSP	.	.	38	97	IR833-6-2
99	DSP	.	.	38	98	IR833-6-2-1
100	DSP	.	.	38	99	IR833-6-2-1-1
101	C3W	100	94	.	.	IR2061
102	BC	85	90	.	.	IR2146
103	DSP	.	.	91	91	IR2006-P3
104	DSP	.	.	91	103	IR2006-P3-31
105	DSP	.	.	91	104	IR2006-P3-31-3
106	DSP	.	.	102	102	IR2146-68
107	DSP	.	.	102	106	IR2146-68-1
108	C2W	105	107	.	.	IR5236
109	DSP	.	.	101	101	IR2061-465
110	DSP	.	.	101	109	IR2061-465-1
111	DSP	.	.	101	110	IR2061-465-1-4
112	DSP	.	.	96	96	IR2055-475
113	DSP	.	.	96	112	IR2055-475-2
114	C2W	111	113	.	.	IR5338
115	C4W	108	114	.	.	IR5657
116	DSP	.	.	115	115	IR5657-33
117	DSP	.	.	115	116	IR5657-33-2
118	DSP	.	.	115	117	IR5657-33-2-1
119	DSP	.	.	101	110	IR2061-465-1-5
120	DSP	.	.	101	119	IR2061-465-1-5-5
121	C2W	118	120	.	.	IR18348
122	DSP	.	.	121	121	IR18348-36
123	DSP	.	.	121	122	IR18348-36-3
124	DSP	.	.	121	123	IR18348-36-3-3
125	VCR			121	124	Cultivar release as IR 64

DISCUSSION AND CONCLUSIONS

The conceptual model for the ICIS GMS balances between allowing for a broad range of reproductive systems and breeding processes while unambiguously identifying the generative, derivative, and maintenance processes used in crop improvement. The model documents historical genealogies, records in crop improvement, management of germplasm stocks, management of research-oriented plant stocks, and management of production-based seed stocks. While the obvious use of the GMS is to archive historical genealogies, the flexibility of the GMS, especially for documenting derivative processes, means that the ICIS GMS can be used for routine management of breeding records.

That the GMS model provides a generic solution is evidenced by the existence of ICIS GMS databases for over 20 crops including rice, wheat (*Triticum aestivum* L. and *T. durum* Desf.), barley (*Hordeum vulgare* L.), maize, cowpea [*Vigna unguiculata* (L.) Walp.], chickpea (*Cicer arietenum* L.), common bean (*Phaseolus vulgaris* L.), lesquerella [*Lesquerella fendleri* (Gray) S. Wats.], witloof chicory (*Cichorium intybus* L.), potato, sweet potato, and taro. The data storage required to identify each genetic entity with a unique GID is not excessive. Large sets of breeding records are readily managed on a personal computer. The International Rice Information System, a full implementation of ICIS, currently holds about 2.6 million unique GIDs (for germplasm accessions, crosses, populations, and lines) and requires about 900 megabytes

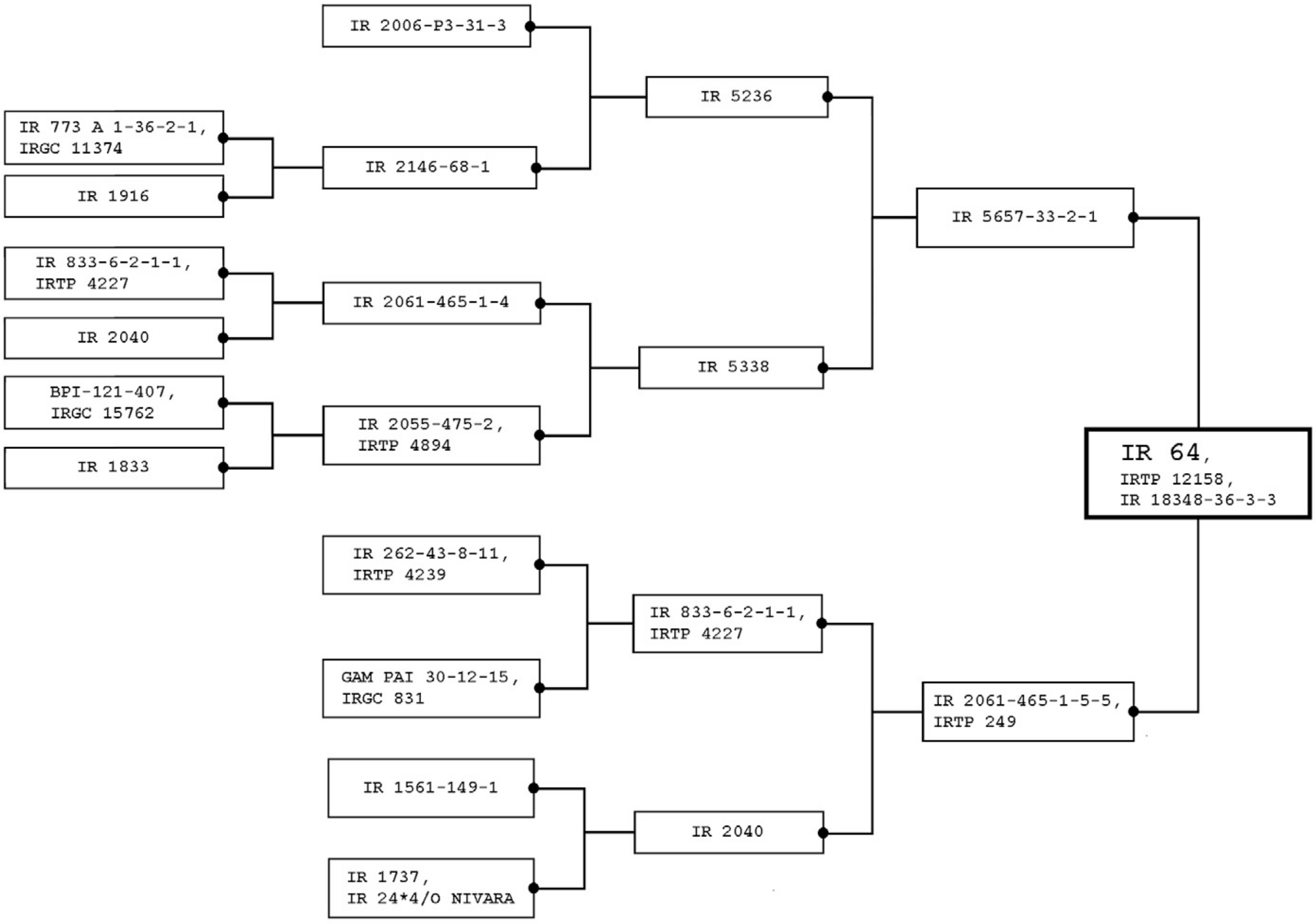

Figure 2. Pedigree of rice line IR 64 represented as a dendrogram.

of storage space. The International Wheat Information System has 5.1 million GIDs and requires 1600 megabytes. Both databases can be accommodated in Microsoft Access (Microsoft Corp., Redmond, WA) databases, and it is possible to use MySQL (Sun Microsystems, Inc., Santa Clara, CA) to store larger datasets.

Regardless of whether one is establishing an implementation for a crop not previously established in ICIS or creating a local implementation, organizing data for loading into GMS requires care. Difficulties in loading data occur with historical data where descriptions are ambiguous on key points. Examples include identification of the parent used as the male and female, particularly in backcrossing, determining whether parents in complex crosses were fixed lines, and deciding whether mass selection should be considered to be among full or half sibs. We emphasize, however, that if necessary, partial records may be loaded using the methods for storing historical pedigree data with incomplete information. Fortunately, software tools can automate much of this process, including loading of external data from existing databases. The biggest challenges to date have been reconciling diverse conventions for representing selection and derivation histories, especially when researchers have inadvertently used different codes to identify the same germplasm, methods, or test environments.

ICIS implementations are available for the crops listed above. Generic tools that are available as part of ICIS include tools for preparing field books, printing labels for seed packets or field plots, displaying genealogical trees, and estimating COPs. Individual users maintain their breeding records in local installations of ICIS and have the option of sharing their records through an update process. Formal user training for ICIS has been provided annually since 1999, and various tutorials are included within the main ICIS documentation provided with the software (www.icis.cgiar.org/icis; accessed 18 Feb. 2009; verified 3 Sept. 2009). A web interface is available for simple queries from a given crop implementation (e.g., the link to IRIS at www.iris.irri.org; accessed 18 Feb. 2009; verified 3 Sept. 2009). The GMS model also enables links to tools for managing phenotypic, molecular, and environmental data. Ongoing efforts seek to improve usability through a web-based, workflow-oriented breeding platform (http://beta.irri.org/seeds/; accessed 18 Feb. 2009; verified 3 Sept. 2009). ICIS requires Microsoft Windows XP or Vista. For further information on ICIS, contact the corresponding author or the ICIS web site.

Table 7. Application of the ICIS GMS model to a breeding scheme for a potato cultivar described by Kirkham and Wilson (1984) and Kirkham (1999). Method codes are as defined in Table 5. Numbers under female parent, male parent, derivative group, and source of derivative group are germplasm identifiers. BID in the Breeder's name column indicates Breeder Identifier (unknown from the cultivar description). A value of "p" is used to indicate that in the complete database, an actual GID number would be provided. Cells containing a period (".") would have a null value in the database.

Germplasm identifier	Method code	No. of progenitors	Female parent	Male parent	Derivative group	Source of derivative group	Breeder's name
301	NCI	.	.	.	p	p	Orion
302	NCI	.	.	.	p	p	V28-12
303	NCI	.	.	.	p	p	Katahdin
304	NCI	.	.	.	p	p	Catriona
401	P2W	2	303	304	.	.	Katahdin/Catriona
421	CDM	.	.	.	401	401	Unknown
431	VCF	.	.	.	401	421	Glen Ilam
441	NCI	.	.	.	401	431	Glen Ilam
501	P2W	2	302	441	.	.	V28_12/Glen Ilam
521	CDM	.	.	.	501	501	60_7_20
531	VCF	.	.	.	501	521	60_7_20
541	NCI	.	.	.	501	531	60_7_20
601	P2W	2	541	301	.	.	Orion/60_7_20
621	CDM	.	.	.	601	601	71-18-4
631	VCF	.	.	.	601	621	Tarago

Acknowledgments

The authors express their special gratitude to software developers and database managers at CIMMYT, IRRI, and elsewhere who have supported ICIS and its predecessors. We also acknowledge the constructive feedback of numerous plant breeders and genebank curators who have commented on the GMS model. Ongoing ICIS development is partially supported IRRI, CIMMYT, Nunhems B.V., Bayer Crop Science, Generation Challenge Program, AAFC SAPRC Swift Current, and Quality Wheat CRC Australia (project completed), GRDC, ACPFG, and University of Queensland in Australia. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture or any other institutions associated with the authors.

References

Bruskiewich, R.M., A.B. Cosico, W. Eusebio, A.M. Portugal, L.M. Ramos, M.T. Reyes, M.A.B. Sallan, V.J.M. Ulat, X. Wang, K.L. McNally, R. Sackville Hamilton, and C.G. McLaren. 2003. Linking genotype to phenotype: The International Rice Information System (IRIS). Bioinformatics 19:63–65.

Cane, K., P.J. Sharp, H.A. Eagles, R.F. Eastwood, G.J. Hollamby, H. Kuchel, M. Lu, and P.J. Martin. 2008. The effects on grain quality traits of a grain serpin protein and the VPM1 segment in southern Australian wheat breeding. Aust. J. Agric. Res. 59:883–890.

Cox, T.S., J.P. Murphy, and D.M. Rodgers. 1986. Changes in genetic diversity in the red winter wheat regions of the United States. Proc. Natl. Acad. Sci. USA 83:5583–5586.

Crossa, J., J. Burgueno, P.L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnamachari. 2006. Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci. 46:1722–1733.

Crossa, J., J. Burgueno, S. Dreisigacker, M. Vargas, S.A. Herrera-Foessel, M. Lillemo, R.P. Singh, R. Trethowan, M. Warburton, J. Franco, M. Reynolds, J. Crouch, and R. Ortiz. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 107:1889–1913.

Fox, P.N., R.I. Magaña, C. Lopez, H. Sanchez, R. Herrera, V. Vicarte, J.W. White, B. Skovmand, and M.C. Mackay. 1997. International Wheat Information System (IWIS), Version 2. [CD-ROM]. CIMMYT, Mexico, D.F.

Fox, P.N., and B. Skovmand. 1996. The International Crop Information System (ICIS) connects genebank to breeder to farmer's field. *In* M. Cooper and G.L. Hammer (ed.) Plant adaptation and crop improvement. CAB International, Wallingford, Oxford, UK.

Gingle, A.R., H. Yang, P.W. Chee, O.L. May, J. Rong, D.T. Bowman, E.L. Lubbers, J.L. Day, and A.H. Paterson. 2006. An integrated web resource for cotton. Crop Sci. 46:1998–2007.

ICIS. 2008a. TDM ontologies and controlled vocabularies. Available at www.cropinfo.org/icis/index.php/TDM_Ontologies_and_Controlled_Vocabularies_5.4 (verified 24 Aug. 2009).

ICIS. 2008b. TDM GMS overview 5.4 Available at cropwiki.irri.org/icis/index.php/TDM_GMS_Overview_5.4 (verified 24 Aug. 2009).

Jordan, D.R., Y.Z. Tao, I.D. Godwin, R.G. Henzell, M. Cooper, and C.L. McIntyre. 2005. Comparison of identity by descent and identity by state for detecting genetic regions under selection in a sorghum pedigree breeding program. Mol. Breed. 14:441–454.

Kirkham, R. 1999. Potato variety—Tarago. Agriculture Notes AG0079. State of Victoria, Dep. of Primary Industries, Melbourne.

Kirkham, R., and G. Wilson. 1984. Tarago: A crisping potato released in South Eastern Australia. Am. J. Potato Res. 61:331–334.

Khush, G.S., and P.S. Virk. 2005. IR varieties and their impact. Int. Rice Res. Inst., Los Baños, Philippines.

McLaren, C.G., R.M. Bruskiewich, A.M. Portugal, and A.B. Cosico. 2005. The International Rice Information System. A platform for meta-analysis of rice crop data. Plant Physiol. 139:637–642.

Parisseaux, B., and R. Bernardo. 2004. In silico mapping of quantitative trait loci in maize. Theoret. Appl. Genet. 109:508–514.

Souza, E., P.N. Fox, and B. Scovmand. 1998. Parentage analysis of International Spring Wheat Yield Nurseries 17 to 27. Crop Sci. 38:337–341.

Stich, B., J. Mohring, H.-P. Piepho, M. Heckenberger, E.S. Buckler, and A.E.R. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754.

Tinker, N.A., and J.K. Deyl. 2005. A curated Internet database of oat pedigrees. Crop Sci. 45:2269–2272.

van Berloo, R., and R.C.B. Hutten. 2005. Peditree: Pedigree database analysis and visualization for breeding and science. J. Hered. 96:465–468.

White, J.W., D.A. Dierig, P. Tomasi, A. Salywon, and D. Nath. 2007. Harnessing information technologies for more efficient crop development. p. 8–18. *In* J. Janick and A. Whipkey (ed.) Issues in new crops and new uses. ASHS Press, Alexandria, VA.

White, J.W., O.V. Voysest, and G. Serrano. 1990. The CIAT bean database. Annu. Rep. Bean Improv. Coop. 33:100–101.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56:330–338.